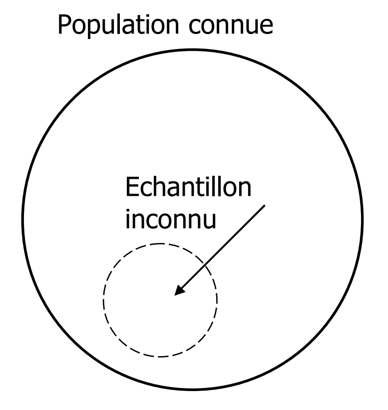
BIOSTATISTIQUE II :
2ème partie : STATISTIQUE INDUCTIVE

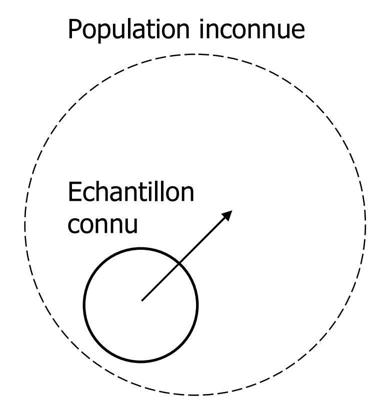
CHAPITRE 1
ECHANTILLONNAGE ET ESTIMATION

I. GENERALITES ET DEFINITIONS

On considère une population sur laquelle on dispose d'informations concernant un paramètre relatif à un certain caractère. L'échantillonnage consiste à passer de la population totale à un échantillon provenant de cette population, c'est à dire à déduire, à partir des informations sur la population, des informations concernant le paramètre sur l'échantillon.
On considère, cette fois, un échantillon sur lequel on dispose d'informations concernant un paramètre relatif à un certain caractère. L'estimation consiste à passer de l'échantillon à la population, c'est à dire à induire, à partir des résultats observés sur l'échantillon, des résultats concernant la population.



Echantillonnage
Déduction du général au particulier



Estimation
Induction du particulier au général

Dans ce qui suit, on traite le problème d'échantillonnage avant de passer à la notion d'estimation.

II. DISTRIBUTION D'ECHANTILLONNAGE ET INTERVALLE DE CONFIANCE D'UNE MOYENNE

A. CAS DES GRANDS ECHANTILLONS ($n \geq 30$)

1. Distribution d'échantillonnage d'une moyenne

On considère une population nombreuse de moyenne M et d'écart-type σ_p relatif à un caractère quantitatif. Si on prélève au hasard k échantillons de même taille n par exemple, on constate que les moyennes $m_1, m_2, \dots, m_k$ de ces k échantillons font apparaître des différences, parfois importantes, dues **aux fluctuations d'échantillonnage**. On désigne par $\bar{X}$, la variable aléatoire qui peut prendre pour valeur la moyenne d'un échantillon prélevé au hasard de la population. $\bar{X}$ est appelée **moyenne d'échantillonnage**.
On détermine la loi de probabilités de $\bar{X}$ appelée **distribution d'échantillonnage de la moyenne**

D'après le théorème central limite, on démontre que :

$$\bar{X} \to N\left(M, \frac{\sigma_p^2}{n}\right) \quad , \quad E(\bar{X}) = M \quad , \quad V(\bar{X}) = \frac{\sigma_p^2}{n}$$

L'intervalle :

$$\boldsymbol{IP(\bar{X}) = \left[M - \ t_\alpha \frac{\sigma_p}{\sqrt{n}}, M + \ t_\alpha \frac{\sigma_p}{\sqrt{n}}\right]}$$

est appelé **intervalle de pari de la moyenne** noté par $\boldsymbol{IP(\bar{X})}$ ou encore **intervalle de fluctuation de la moyenne**. C'est l'intervalle qui contient $\bar{X}$ au risque d'erreur $\boldsymbol{\alpha}$.

$\boldsymbol{1 - \alpha}$ est appelé **seuil de confiance**.
$\boldsymbol{\alpha}$ est appelé **risque d'erreur**.
$\boldsymbol{t_\alpha}$ est une valeur donnée par la table de la loi normale centrée réduite.
En général, on choisit $\boldsymbol{\alpha = 5\%}$ et dans certains cas assez particuliers $\boldsymbol{\alpha = 1\%}$.
D'après les propriétés de la loi normale on a :

$\boldsymbol{\alpha = 5\%}$, $\boldsymbol{t_\alpha = 1,96}$

$\boldsymbol{\alpha = 1\%}$, $\boldsymbol{t_\alpha = 2,6}$

Exemple

Une machine est destinée à fabriquer des comprimés de poids moyen de 200 mg avec un écart-type de 10 mg. On extrait au hasard un échantillon de 50 comprimés. Entre quelles limites varie le poids moyen des comprimés de cet échantillon au risque de 5% ?

Solution

Population : $M = 200$ et $\sigma_p = 10$
Echantillon : $n = 50 > 30$
Le poids moyen d'un échantillon varie d'un échantillon à un autre, c'est donc une v.a. que l'on désigne par $\bar{X}$ (la moyenne d'échantillonnage).
Comme $n > 30$ alors

$$\bar{X} \to N\left(M, \frac{\sigma_p^2}{n}\right) \quad , \quad E(\bar{X}) = M \quad , \quad V(\bar{X}) = \frac{\sigma_p^2}{n}$$

L'intervalle de pari de $\bar{X}$ est donc :

$$IP(\bar{X}) = \left[M - t_\alpha \frac{\sigma_p}{\sqrt{n}}, M + t_\alpha \frac{\sigma_p}{\sqrt{n}}\right]$$

au risque α donné.
A.N. : Pour $\alpha = 5\%$, $t_\alpha = 1,96$

$IP(\bar{X}) = [197,22 \, , 202,77]$ au risque $\alpha = 5\%$

Le poids moyen d'un échantillon de 50 comprimés est compris entre 197,22 et 202,77 avec un risque de 5% de se tromper.

2. Intervalle de confiance d'une moyenne

Soit à étudier dans une population un certain caractère quantitatif. Désignons par M la moyenne et σ_p l'écart-type du caractère étudié (M et σ_p sont inconnus). On prélève au hasard un échantillon de taille n et on en détermine la moyenne m et l'écrat-type σ_e. Le problème qui se pose est d'estimer la moyenne M de la population à partir de n, m et σ_e, c'est à dire de trouver un intervalle dans lequel se trouve la moyenne de la population M.

L'intervalle :

$$\boldsymbol{IC(M) = \left[m - t_\alpha \frac{\sigma_e}{\sqrt{n-1}}, m + t_\alpha \frac{\sigma_e}{\sqrt{n-1}}\right]}$$

est appelé **intervalle de confiance de la moyenne** noté par $\boldsymbol{IC(M)}$. C'est l'intervalle qui contient la moyenne M de la population au risque d'erreur $\boldsymbol{\alpha}$. On choisit généralement $\boldsymbol{\alpha = 5\%}$ ou $\boldsymbol{\alpha = 1\%}$.

Pour $\boldsymbol{\alpha = 5\%}$, $\boldsymbol{t_\alpha = 1,96}$
Pour $\boldsymbol{\alpha = 1\%}$, $\boldsymbol{t_\alpha = 2,6}$

Remarque :

σ_p étant inconnu et on démontre que, lorsque $n \geq 30$, la variance de la population σ_p^2 est estimée par $\frac{n}{n-1}\sigma_e^2$. En d'autre terme, lorsque $\boldsymbol{n \geq 30}$ on a :

$$\boldsymbol{\sigma_p^2 \approx \frac{n}{n-1}\sigma_e^2}$$

Donc s^2 est la variance estimée de la population et on a :

$$\boldsymbol{\sigma_p^2 \approx s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2}$$

Exemple

Dans une population de personnes, on extrait au hasard un échantillon de taille 40 dont le poids moyen est de 70 Kg et l'écart-type de 15,4 Kg. Quel est, au risque de 5%, l'intervalle de confiance du poids moyen de la population ?

Solution

Echantillon : $n = 40 > 30 \quad m = 70$ et $\sigma_e = 15{,}4$

Désignons par M le poids moyen de la population à estimer.

L'intervalle de confiance de M est donc :

$$IC(M) = \left[m - t_\alpha \frac{\sigma_e}{\sqrt{n-1}}, m + t_\alpha \frac{\sigma_e}{\sqrt{n-1}}\right]$$

au risque α donné.

A.N. : Pour $\alpha = 5\%$, $t_\alpha = 1{,}96$

$$IC(M) = [65{,}16\,, 74{,}83] \text{ au risque } \alpha = 5\%$$

Ceci veut dire qu'il y a 95% de chances pour que l'intervalle de confiance $[65{,}16\,, 74{,}83]$ contienne le poids moyen M de la population.

3. Précision de l'estimation

Il convient de remarquer que la précision de l'estimation est d'autant meilleure que la taille de l'échantillon est assez grande car la longueur de l'intervalle de confiance diminue quand n croît.

On a :

$$M = m \pm t_\alpha \frac{\sigma_e}{\sqrt{n-1}}$$

La précision de l'estimation est donc :

$$\boxed{\boldsymbol{h = t_\alpha \frac{\sigma_e}{\sqrt{n-1}}}}$$

pour un risque α donné.

Dans l'exemple précédent, La précision de l'estimation du poids moyen de la population est :

$$h = 1{,}96 \frac{15{,}4}{\sqrt{39}} = 4{,}83$$

D'autre part, si on diminue le risque α (donc t_α augmente), la longueur de l'intervalle de confiance augmente, par conséquent on perd la précision de l'estimation.

4. Détermination du nombre d'individus nécessaire

Revenons à notre exemple précédent.

Supposons qu'on veuille estimer le poids moyen de la population à 2Kg près. Quel devrait être le nombre minimum d'individus nécessaire pour atteindre cette précision au risque $\alpha = 5\%$?

En d'autres termes, on détermine n' tel que $h = 2$.

On a donc :

$$\boldsymbol{n' = \left(\frac{t_\alpha . \sigma_e}{2}\right)^2 + 1}$$

Au risque $\alpha = 5\%$, $t_\alpha = 1{,}96$ et $\sigma_e = 15{,}4 \Rightarrow n' = 228{,}76 \approx 229$

La taille minimale de l'échantillon devrait être égale à 229 pour atteindre la précision désirée (2 Kg).

Plus généralement, pour un risque α et une précision désirée h on a :

$$\boldsymbol{n} = \left(\frac{\boldsymbol{t_\alpha . \sigma_e}}{\boldsymbol{h}}\right)^{\mathbf{2}} + \mathbf{1}$$

Lors de l'estimation de la moyenne M de la population, il est possible de déterminer le nombre minimum d'individus nécessaire à condition :

1. de fixer à l'avance une précision h et un risque α,
2. de connaître l'écart-type d'un échantillon préalablement étudié (σ_e).

B. CAS DES PETITS ECHANTILLONS ($n < 30$)

1. Distribution d'échantillonnage d'une moyenne

Lorsque $n < 30$, la moyenne d'échantillonnage $\bar{X}$ ne suit pas en général une loi normale sauf si le caractère étudié dans la population suit une loi normale. Dans ce qui suivra, on suppose que l'hypothèse de normalité du caractère étudié est vérifiée (d'ailleurs, cette hypothèse est très souvent réalisée en médecine et en biologie).

Comme dans le cas des grands échantillons on a :

$$\bar{X} \to N\left(M, \frac{\sigma_p^2}{n}\right) \quad , \quad E(\bar{X}) = M \quad , \quad V(\bar{X}) = \frac{\sigma_p^2}{n}$$

L'**intervalle de pari de la moyenne** noté par $\boldsymbol{IP}(\bar{\boldsymbol{X}})$ ou encore **intervalle de fluctuation de la moyenne** est :

$$\boldsymbol{IP}(\bar{\boldsymbol{X}}) = \left[\boldsymbol{M} - \boldsymbol{t_\alpha} \frac{\boldsymbol{\sigma_p}}{\sqrt{\boldsymbol{n}}}, \boldsymbol{M} + \boldsymbol{t_\alpha} \frac{\boldsymbol{\sigma_p}}{\sqrt{\boldsymbol{n}}}\right]$$

En général, on choisit $\boldsymbol{\alpha} = \mathbf{5}\%$ et dans certains cas assez particuliers $\boldsymbol{\alpha} = \mathbf{1}\%$.
Pour $\boldsymbol{\alpha} = \mathbf{5}\%$, $\boldsymbol{t_\alpha} = \mathbf{1,96}$
Pour $\boldsymbol{\alpha} = \mathbf{1}\%$, $\boldsymbol{t_\alpha} = \mathbf{2,6}$

2. Intervalle de confiance d'une moyenne

Le problème d'estimation est analogue à celui posé précédemment. On connaît n, m et σ_e.

L'**intervalle de confiance de la moyenne** noté par $\boldsymbol{IC}(\boldsymbol{M})$ est :

$$\boldsymbol{IC}(\boldsymbol{M}) = \left[\boldsymbol{m} - \boldsymbol{t_\alpha^*} \frac{\boldsymbol{\sigma_e}}{\sqrt{\boldsymbol{n}}}, \boldsymbol{m} + \boldsymbol{t_\alpha^*} \frac{\boldsymbol{\sigma_e}}{\sqrt{\boldsymbol{n} - \mathbf{1}}}\right]$$

où $\boldsymbol{t_\alpha^*}$ est la valeur donnée par la table de STUDENT-FISHER en fonction du risque α=5% ou α=1% et le nombre de degré de liberté $\nu = n - 1$.

Condition d'utilisation : Cette formule nécessite la condition de normalité du caractère.

Exemple

Un dosage de sucre dans une solution effectué sur 8 prélèvements provenant d'une même population a donné les résultats suivants exprimés en g/l.

19,5 19,7 19,8 20,2 20,2 20,3 20,4 20,8

1- Calculer la moyenne et l'écart-type de cette distribution.
2- Quel est l'intervalle de confiance de la moyenne au risque de 5% ?

Solution

1- Calcul de la moyenne et de l'écart-type

$$m = \frac{1}{8}\sum_{i=1}^{8} x_i = 20{,}11$$

$$\sigma_e = \sqrt{\frac{1}{8}\sum_{i=1}^{8} (x_i - m)^2} = 0{,}395$$

2- Désignons par M le dosage moyen du sucre de la population à estimer. En supposant que le dosage du sucre est distribué dans la population selon une loi normale, l'intervalle de confiance de la moyenne M est donc :

$$IC(M) = \left[m - t_\alpha^* \frac{\sigma_e}{\sqrt{n-1}}, m + t_\alpha^* \frac{\sigma_e}{\sqrt{n-1}}\right]$$

au risque α.

A.N. : Au risque $\alpha = 5\%$ avec $\nu = 8 - 1 = 7$ la table de STUDENT-FISHER nous donne $t_\alpha^* = 2{,}365$.

$$I.C(M) = [19{,}75\ , 20{,}46] \text{ au risque } \alpha = 5\%$$

III. DISTRIBUTION D'ECHANTILLONNAGE ET INTERVALLE DE CONFIANCE D'UNE PROPORTION

1. Distribution d'échantillonnage d'une proportion

Soit p la proportion d'individus porteurs d'un caractère dans une population nombreuse. On extrait au hasard un échantillon. Soit la v.a. Y, la proportion d'individus qui portent le caractère. Y est appelée **proportion d'échantillonnage**.

On détermine la loi de probabilité de Y appelée **distribution d'échantillonnage de la proportion.**

Lorsque $n.p \geq 5$, on démontre que :

$$Y \to N\left(p, \frac{p.q}{n}\right) \quad , \quad E(Y) = p \quad , \quad V(Y) = \frac{p.q}{n} \quad , \quad q = 1 - p$$

L'**intervalle de pari de la proportion** noté par $\boldsymbol{IP(Y)}$ ou encore **intervalle de fluctuation de la proportion** est :

$$\boldsymbol{IP(Y) = \left[p - t_\alpha\sqrt{\frac{p.q}{n}}, p + t_\alpha\sqrt{\frac{p.q}{n}}\right]}$$

En général, on choisit $\boldsymbol{\alpha = 5}\%$ et dans certains cas assez particuliers $\boldsymbol{\alpha = 1}\%$.

Pour $\boldsymbol{\alpha = 5}\%$, $\boldsymbol{t_\alpha = 1,96}$

Pour $\boldsymbol{\alpha = 1}\%$, $\boldsymbol{t_\alpha = 2,6}$

Exemple

Chez une race de souris on a trouvé que la présence de cancers spontanés est de 25%. Dans quel intervalle, au risque de 5%, est situé le pourcentage de cancers pour un échantillon de 100 souris ?

Solution

Population : $p = 0{,}25$

Echantillon : $n = 100$

Le pourcentage de cancers varie d'un échantillon à un autre, c'est donc une v.a. que l'on désigne par Y (la proportion d'échantillonnage).

L'intervalle de pari de Y, correspondant au risque α donné, est donc :

$$IP(Y) = \left[p - t_\alpha\sqrt{\frac{p.q}{n}}, p + t_\alpha\sqrt{\frac{p.q}{n}}\right]$$

La condition de validité est remplie : $n.p = 100.0{,}25 = 25 > 5$.

A.N. : Pour $\alpha = 5\%$, $t_\alpha = 1{,}96$

$$IP(Y) = [0{,}16{,}0{,}33] \text{ au risque } \alpha = 5\%$$

La proportion de cancers pour un échantillon de 100 souris est située dans l'intervalle [0,16, 0,33] au risque de 5%.